Link_Char_Model

Naomi Giertych

6/18/2021

What characteristics make a link appealing to click on?

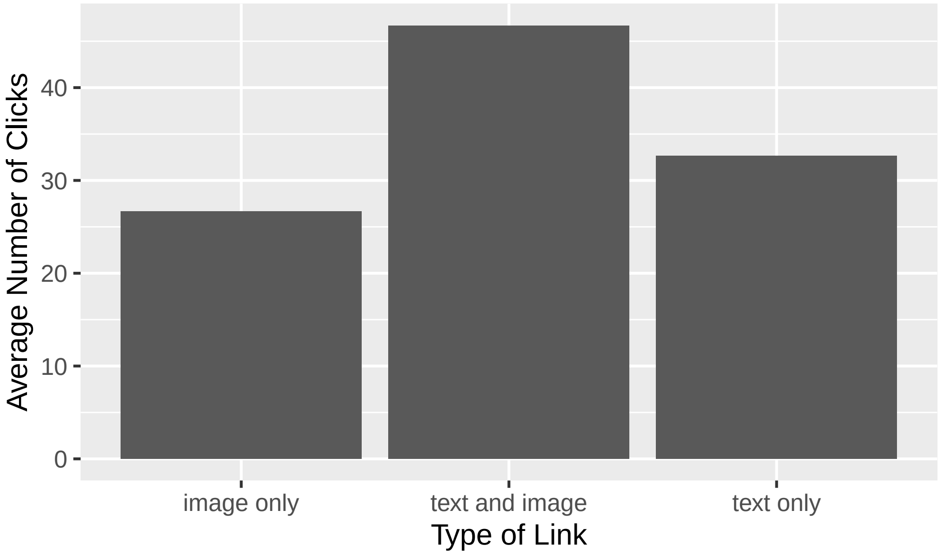
In this report, we investigate the characteristics of a links that make it more likely to be clicked on. We focus on newsletters from January 2019 to December 2020. Before diving into these characteristics, we give a brief description of how the data used was obtained.

Data Description

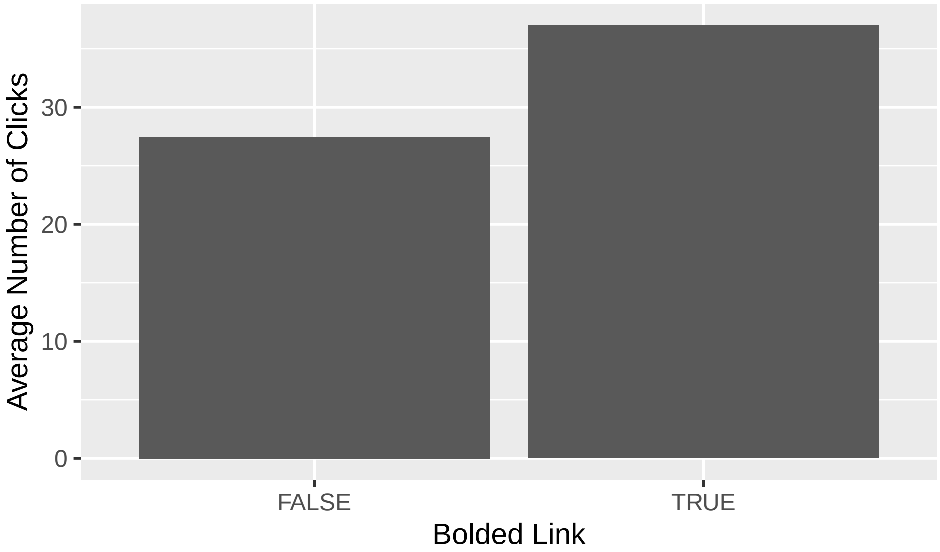
The data comes from a few sources: the CSV and raw text files generated from iContact and the HTML source code from the links in each of the newsletters. From the CSV files, we determine the unique number of times a link was clicked on. We define a unique click to be a unique combination of subscriber ID, newsletter date, and link; in other words, if a subscriber clicked on the same link from the same newsletter, we do not count that click. We also identify the time of day the newsletter was sent and whether it was before or after the COVID-19 pandemic was declared (03/20/20) from the CSV files. The raw text files are used to get an approximation of how far down the newletter the link is, e.g. a link that is about half-way down the newsletter would be estimated around 50%. Finally, we obtain style characteristics and whether the link was an image or had an image associated with it from the HTML source code.

Data Exploration

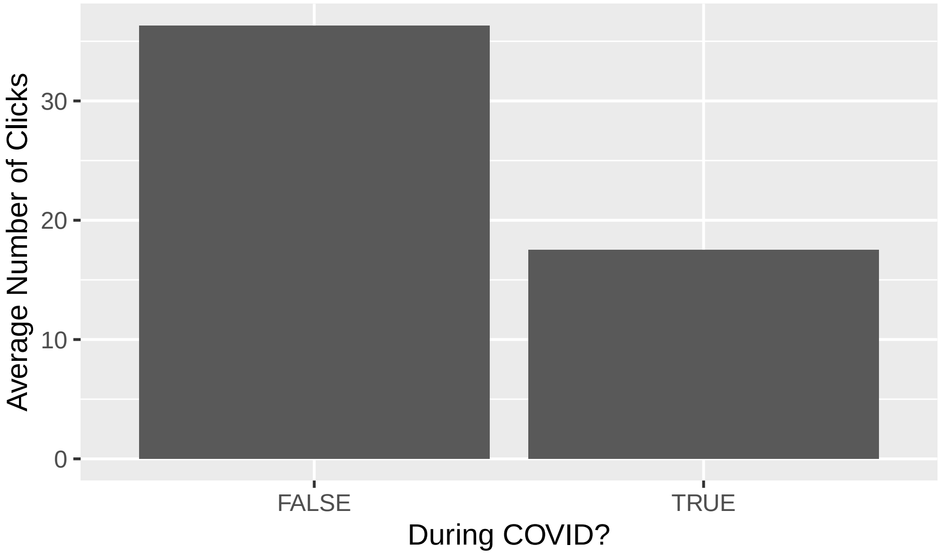
Before fitting any models to the data, we explore how the number of clicks a link received depended on the variables mentioned above. We define the average number of times a link was clicked as the number of times the link was uniquely clicked, defined above, divided by the number of newsletters the link appeared in. It is important to note that in doing this, we do not control for how many times a link was used within the same newsletter. For each of the categorical variables, we graph the category and the average number of tiems a link was clicked below.



In the bar plot above, the label ‘image only’ refers to links that only had a picture associated with it, ‘text only’ refers to links with only associated text, and ‘text and image’ refers to links with both an image and associated text. Based on the bar plot, it appears that a combination of text and pictures encourages people to click on a link.

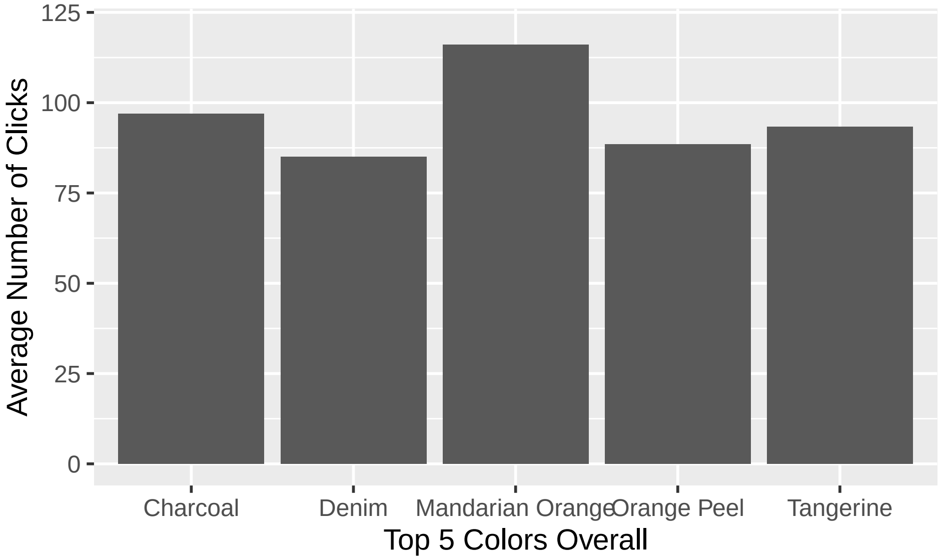


Based on the bar plot above, it appears the bolding the text associated with the link also increases the chance that someone clicks on it.



Based on the bar plot above, it seems COVID impacted how much subscribers choose to interact with the newsletters. This is not all that surprising given the challenges everyone was facing during the pandemic.

In the next two bar plots, we focus on the top five text color choices across all newsletters and the top five text color choices that appeared in more than one newsletter.



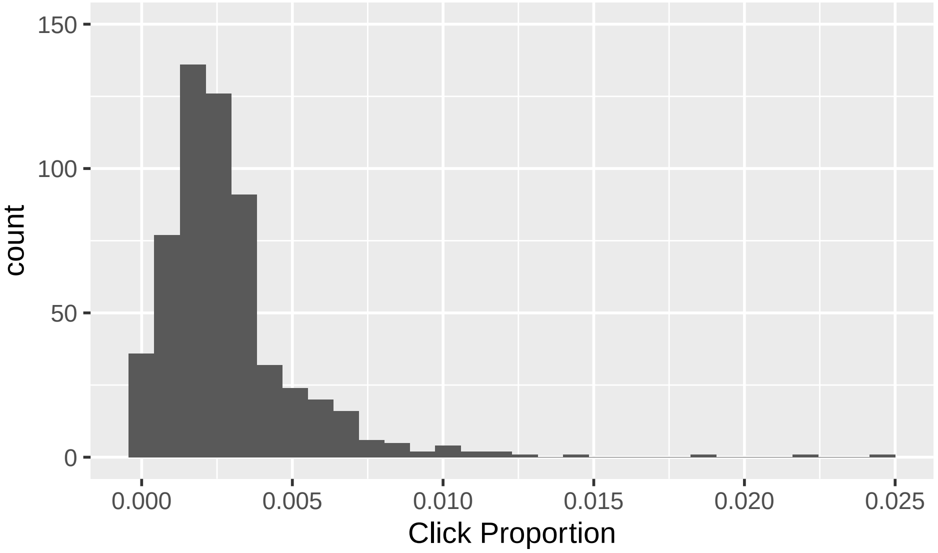
Any color of orange seems to grab people's attention! Mandarian Orange only appeared in the newletter promoting the Remote Volunteer Project: DIY Family Essentials Kits opportunity so it is tempting to think the large number of clicks this color received may have more to do with the highly-relatable project. However, this project was advertised in four different newsletters using links colored as cinnabar and falu red (both are different tints of red) and these other newsletters had less than 85 clicks each. While there are more factors at play than just the link color, the fact that the newsletters advertising the same opportunity in red got fewer clicks suggests that a text color of orange is more impactful.

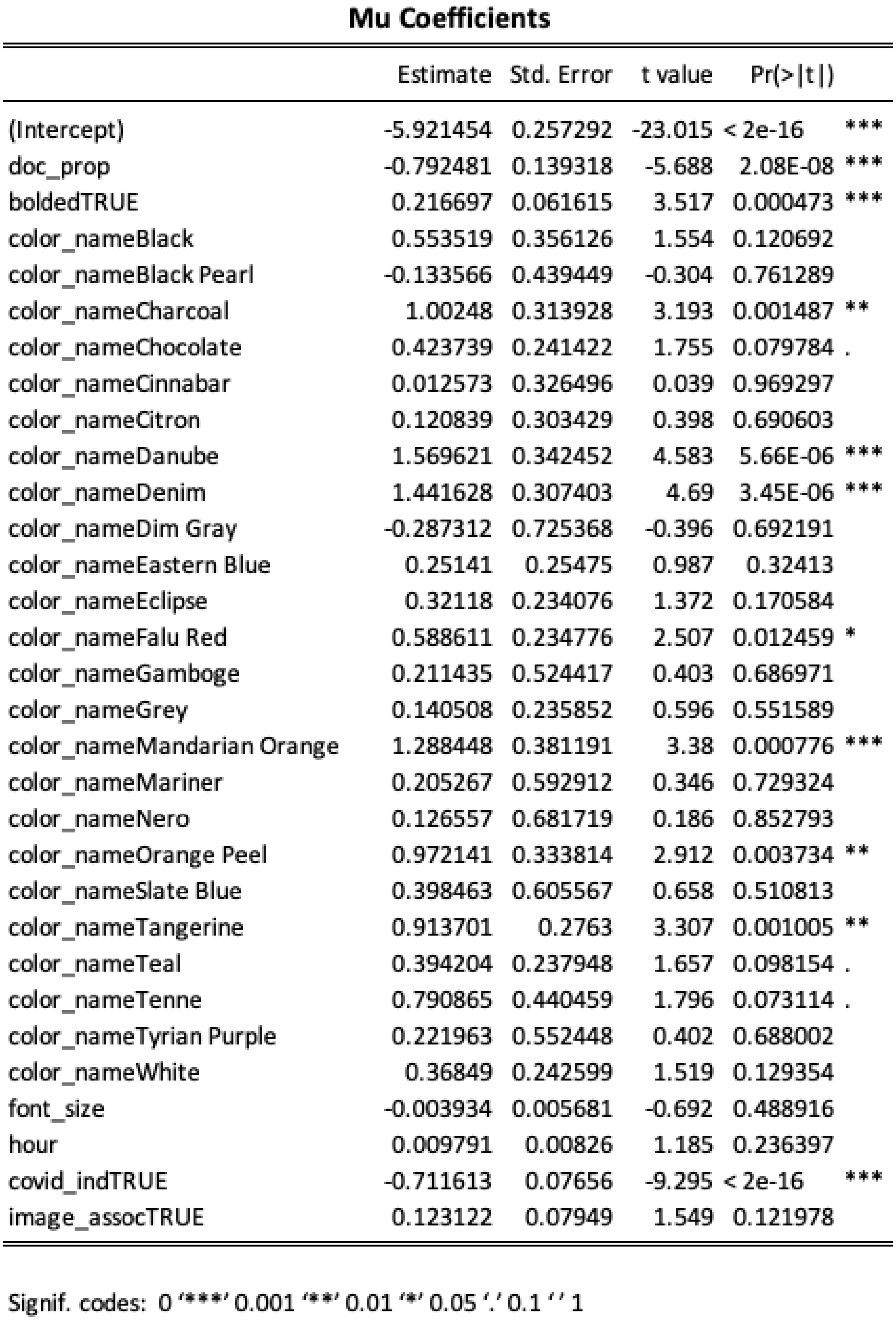
Model Fitting

We attempted to fit a linear regression, a Poisson regression, and a Negative Binomial regression to the click data with no success. However, we found that if we divided the number of clicks by the number of subscribers each newsletter was sent to, a zero-inflated beta regression worked quite well. The beta regression allows us to model proportion data (data that's bounded between zero and one, non-inclusive); the "zero-inflated" in the name refers to extending the beta regression to include observations with a value of zero. The zero-inflated beta regression fits three parameters: mu, sigma, and nu. The mu variable corresponds to the mean of the click proportion (relative to the number of subscribers) and is modeled in a similar manner to simple linear regression.

The variables in our model are the following: doc_prop, bolded, color, font_size, hour, covid_ind, imag_assoc. "doc_prop" is the proportion down the document a link is; in other words, a link that is about halfway down a newsletter will be about 50%. "bolded" indicates whether a link was bolded. "color" is the color of the link as determined by https://www.color-blindness.com/color-name-hue/.

Below we give a histogram of click proportion and the fitted model parameters for mu. From the table below, we see that where the link is in the document, whether the link is bolded, and the color of the link make a statistically signficant difference on whether the link is clicked or not. Additionally, we see the top five colors shown above are also statistically signficant as well as Charcoal and Falu Red.





Mu Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-5.921454	0.257292	-23.015	< 2e-16	***
doc_prop	-0.792481	0.139318	-5.688	2.08E-08	***
boldedTRUE	0.216697	0.061615	3.517	0.000473	***
color_nameBlack	0.553519	0.356126	1.554	0.120692	
color_nameBlack Pearl	-0.133566	0.439449	-0.304	0.761289	
color_nameCharcoal	1.00248	0.313928	3.193	0.001487	**
color_nameChocolate	0.423739	0.241422	1.755	0.079784	.
color_nameCinnabar	0.012573	0.326496	0.039	0.969297	
color_nameCitron	0.120839	0.303429	0.398	0.690603	
color_nameDanube	1.569621	0.342452	4.583	5.66E-06	***
color_nameDenim	1.441628	0.307403	4.69	3.45E-06	***
color_nameDim Gray	-0.287312	0.725368	-0.396	0.692191	
color_nameEastern Blue	0.25141	0.25475	0.987	0.32413	
color_nameEclipse	0.32118	0.234076	1.372	0.170584	
color_nameFalu Red	0.588611	0.234776	2.507	0.012459	*
color_nameGamboge	0.211435	0.524417	0.403	0.686971	
color_nameGrey	0.140508	0.235852	0.596	0.551589	
color_nameMandarian Orange	1.288448	0.381191	3.38	0.000776	***
color_nameMariner	0.205267	0.592912	0.346	0.729324	
color_nameNero	0.126557	0.681719	0.186	0.852793	
color_nameOrange Peel	0.972141	0.333814	2.912	0.003734	**
color_nameSlate Blue	0.398463	0.605567	0.658	0.510813	
color_nameTangerine	0.913701	0.2763	3.307	0.001005	**
color_nameTeal	0.394204	0.237948	1.657	0.098154	.
color_nameTenne	0.790865	0.440459	1.796	0.073114	.
color_nameTyrian Purple	0.221963	0.552448	0.402	0.688002	
color_nameWhite	0.36849	0.242599	1.519	0.129354	
font_size	-0.003934	0.005681	-0.692	0.488916	
hour	0.009791	0.00826	1.185	0.236397	
covid_indTRUE	-0.711613	0.07656	-9.295	< 2e-16	***
image_assocTRUE	0.123122	0.07949	1.549	0.121978	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1